W32–W37 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl305

PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence

Z. R. Li[1,2], H. H. Lin[1], L. Y. Han[1], L. Jiang[1], X. Chen[3] and Y. Z. Chen[1,4,*]

[1]Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543,, [2]College of Chemistry, Sichuan University, Chengdu, 610064, P. R. China, [3]Department of Biotechnology, Zhejiang University, Hangzhou, 310029, P. R. China and [4]Shanghai Center for Bioinformation Technology, Shanghai, 201203, P. R. China

Received December 23, 2005; Revised January 17, 2006; Accepted April 10, 2006

ABSTRACT

Sequence-derived structural and physicochemical features have frequently been used in the development of statistical learning models for predicting proteins and peptides of different structural, functional and interaction profiles. PROFEAT (Protein Features) is a web server for computing commonly-used structural and physicochemical features of proteins and peptides from amino acid sequence. It computes six feature groups composed of ten features that include 51 descriptors and 1447 descriptor values. The computed features include amino acid composition, dipeptide composition, normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, sequence-order-coupling number, quasi-sequence-order descriptors and the composition, transition and distribution of various structural and physicochemical properties. In addition, it can also compute previous autocorrelations descriptors based on user-defined properties. Our computational algorithms were extensively tested and the computed protein features have been used in a number of published works for predicting proteins of functional classes, protein–protein interactions and MHC-binding peptides. PROFEAT is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi

INTRODUCTION

Sequence-derived structural and physicochemical features have frequently been used for predicting protein structural and functional classes (1–5), protein–protein interactions (6–8), subcellular locations (9,10) and peptides of specific properties (11) (J. Cui, L. Y. Han, H. H. Lin, H. L. Zhang, Z. Q. Tang, C. J. Zheng, Z. W. Cao and Y. Z. Chen, manuscript submitted) from their sequence. These features are highly useful for representing and distinguishing proteins or peptides of different structural, functional and interaction profiles, which is essential for the successful application of statistical learning methods in predicting the structural, functional and interaction profiles of proteins and peptides irrespective of sequence similarity (12). While several programs for computing protein structural and physicochemical features have been developed (1,2,6,9–11,13), these are not freely and easily accessible. We introduce PROFEAT, **Pro**tein **Feat**ures, as a freely accessible web-based server for computing the commonly-used structural and physicochemical features of proteins and peptides from amino acid sequence.

WEB SERVER ACCESS

PROFEAT is available at http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi. The sequence of a protein or a peptide, in single-letter code and RAW format, as well as FASTA format, can be input in a window provided. The RAW format is similar to the plain text format except that it removes any white-space and TAB characters, accepts only alphabetic characters and rejects anything else. Multiple sequence entries, in FASTA format, can also be input to facilitate the convenient export of the generated protein features to machine learning methods servers. Illustrative examples for submitting single sequence entry and multiple sequence files to POFEAT and for sending the generated feature vector files to a machine learning server GIST (14) are provided in the on-line manual at the PROFEAT homepage.

An input sequence with less than eight amino acids is not accepted, because functional peptides typically contain more

*To whom correspondence should be addressed. Tel: +65 6516 6877; Fax: +65 6774 6756; Email: yzchen@cz3.nus.edu.sg

than eight amino acids and protein chains are much longer. If an input sequence contains an invalid character, or a non-amino acid letter, or abnormal composition, such as long stretch of the same amino acid covering an entire protein sequence, then a message of 'invalid character . . .' or 'your input sequence is invalid' is displayed. The computed features are divided into six groups each of which has been separately used for protein or peptide studies. Upon submitting a sequence, users are directed to a window shown in Figure 1 for selecting the feature groups to be displayed and the output file format. Three types of file format are provided to support printer-friendly view and the export of the computed features to computational software or servers, such as GIST (14). An index Fi.j.k.l is used to represent the *l*th descriptor value of the *k*th descriptor of the *j*th feature in the *i*th feature group, which serves as an easy reference to the PROFEAT manual provided in the server homepage.

MATERIALS AND METHODS

As shown in Table 1, 10 sets of commonly-used structural and physicochemical features, including 51 descriptors and 1447 descriptor values, are computed by PROFEAT. These features can be divided into six groups each of which has been used as an independent set of features for predicting proteins and peptides of various profiles by using statistical learning methods. The first group includes two features, amino acid composition and dipeptide composition, with 2 descriptors and 420 descriptor values (15–20). Each of the second, third and fourth group contains a different autocorrelation feature: normalized Moreau–Broto autocorrelation (21,22), Moran autocorrelation (23) and Geary autocorrelation (24). Each of these features has 8 descriptors and 240 descriptor values. The fifth group consists of three feature sets: composition, transition and distribution with a total of 21 descriptors and 147 descriptor values (2–6,8,25,26) (J. Cui, L. Y. Han, H. H. Lin, H. L. Zhang, Z. Q. Tang, C. J. Zheng, Z. W. Cao and Y. Z. Chen, manuscript submitted). The sixth group contains two sequence-order feature sets (9–11,27), one is sequence-order-coupling number with 2 descriptors and 60 descriptor values, and the other is quasi-sequence-order with 2 descriptors and 100 descriptor values. Apart from these descriptors, it can also compute previous autocorrelation descriptors based on user-defined properties. The references of the studies that used which of these features are provided in the subsequent discussions.

Amino acid and dipeptide composition are simplistic descriptors of protein sequence features (15), which have been used for predicting protein fold and structural classes (19,20), functional classes (16) and subcellular locations (17,18) at accuracy levels of 72–95%, 83–97% and 79–91%, respectively. Amino acid composition is the fraction of each amino acid type in a sequence: $f(r) = N_r/N$, where $r = 1, 2, 3, \ldots, 20$, N_r is the number of amino acid of type r, and N is the length of the sequence. A total of 20 descriptor values are computed for the 20 types of amino acids. Dipeptide composition is defined as: $fr(r,s) = N_{rs}/(N-1)$, where $r,s = 1, 2, 3, \ldots, 20$, and N_{rs} is the number of dipeptides of amino acid type r and s (16). A total of 400 descriptor values are computed for the 20×20 amino acid combinations.

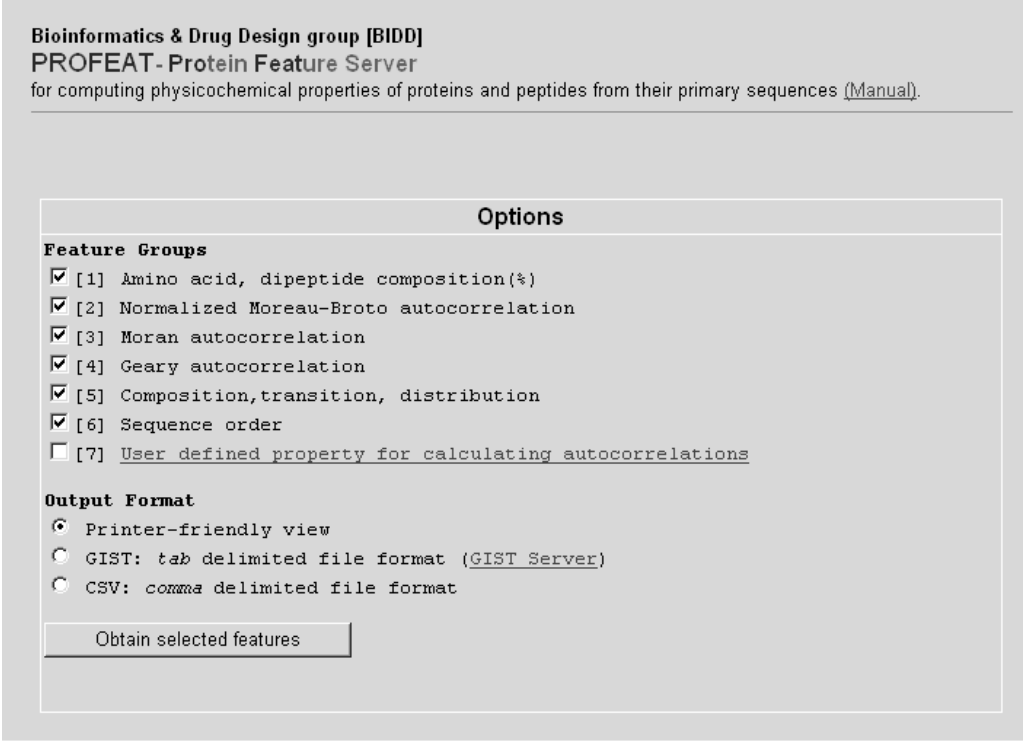

Figure 1. PROFEAT feature-display options window

Table 1. List of structural and physicochemical features of proteins and peptides commonly-used for predicting proteins and peptides of specific properties by using statistical learning methods

Feature group	Feature	Feature index	No. of descriptors	No. of descriptor values
Amino acid, dipeptide composition	Amino acid composition	F1.1	1	20
	Dipeptide composition	F1.2	1	400
Autocorrelation 1	Normalized Moreau–Broto autocorrelation	F2.1	8	240
Autocorrelation 2	Moran autocorrelation	F3.1	8	240
Autocorrelation 3	Geary autocorrelation	F4.1	8	240
Composition, transition, distribution	Composition	F5.1	7	21
	Transition	F5.2	7	21
	Distribution	F5.3	7	105
Sequence-order	Sequence-order-coupling number	F6.1	2	60
	Quasi-sequence-order descriptors	F6.2	2	100

Autocorrelation features describe the level of correlation between two objects (protein or peptide sequences) in terms of their specific structural or physicochemical property (28), which are defined based on the distribution of amino acid properties along the sequence (29). There are eight amino acid properties used for deriving these autocorrelation descriptors. The first is hydrophobicity scale derived from the bulk hydrophobic character for the 20 types of amino acids in 60 protein structures (30). The second is the average flexibility index derived from the statistical average of the B-factors of each type of amino acids in the available protein X-ray crystallographic structures (31). The third is the polarizability parameter computed from the group molar refractivity values originally provided by Hansch *et al*. (32). The fourth is the free energy of amino acid solution in water measured by Hutchins (32). The fifth is the residue accessible surface areas taken from average values from folded proteins (33). The sixth is the amino acid residue volumes measured by Fisher (34). The seventh is the steric parameters derived from the van der Waals raddi of amino acid side-chain atoms (35). The eighth is the relative mutability obtained by multiplying the number of observed mutations by the frequency of occurrence of the individual amino acids (36). Each of these properties is centralized and standardized such that $P_r' = (P_r - \bar{P})/\sigma$, where $\bar{P}$ is the average of the property of the 20 amino acids, $\bar{P}$ and σ are given by:

$$\bar{P} = \frac{\sum_{r=1}^{20} P_r}{20} \quad \mathbf{1}$$

$$\sigma = \sqrt{\frac{1}{20}\sum_{r=1}^{20}(P_r - \bar{P})^2} \quad \mathbf{2}$$

Three different autocorrelation features are computed, each having 8 descriptors and 240 descriptor values. The first is Moreau–Broto autocorrelation $AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d}$ (28,37), which has been used for predicting transmembrane protein types (21) and protein secondary structural contents (22) at accuracy levels of 82–94% and 91–94%, respectively. Here d is the lag of the autocorrelation, P_i and P_{i+d} are the amino acid property at position i and $i+d$, respectively. The normalized Moreau–Broto autocorrelation is defined as: $ATS(d) = AC(d)/(N-d)$ where $d = 1, 2, 3, \ldots, 30$. The second is Moran autocorrelation (38), which has been applied for predicting protein helix contents at an accuracy level of 85% (23), and it is defined as:

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \bar{P})(P_{i+d} - \bar{P})}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad d = 1, 2, 3, \ldots, 30 \quad \mathbf{3}$$

where d and P_i and P_{i+d} are defined above, $\bar{P}$ is the average of P_i, i.e. $\bar{P} = (\sum_{i=1}^{N} P_i)/N$. This algorithm differs from that of Moreau–Broto autocorrelation in the use of property deviations from the average values instead of the property values themselves as the basis for measuring correlations. The third feature is Geary autocorrelation (39), which has been used for analyzing allele frequencies and population structures (24), and it is defined as:

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad d = 1, 2, 3, \ldots, 30 \quad \mathbf{4}$$

where d, $\bar{P}$, P_i and P_{i+d} are defined above. This algorithm differs from the other two algorithms in the use of square-difference of property values instead of vector-product of property values or deviations as the basis for measuring correlations.

Composition, transition and distribution features represent the amino acid distribution patterns of a specific structural or physicochemical property along a protein or peptide sequence (5,25), which have been used for recognition of protein folds (5) and prediction of protein–protein interactions (6,8), protein functional families (2–4,26) and MHC-binding peptides (J. Cui, L. Y. Han, H. H. Lin, H. L. Zhang, Z. Q. Tang, C. J. Zheng, Z. W. Cao and Y. Z. Chen, manuscript submitted) at accuracy levels of 74–100%, 77–81%, 67–99%, 97–99%, respectively. Seven types of physicochemical properties have been used for computing these features. These are hydrophobicity, normalized Van der Waals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility (2,5,25).

These descriptors are computed by the following procedure: the sequence of the amino acids is transformed into a sequence of certain structural or physicochemical properties (attributes) of residues. Twenty amino acids are divided into three groups for each of the seven different attributes based on the main clusters of the amino acid indices of Tomii and Kanehisa (5,40). The reason for dividing amino acids into three groups instead of two or four groups is that, while amino acids can be divided into a minimum of both two and three groups for most attributes, they can only be divided into a minimum of three groups for such attributes as charge (positive, negative and neutral) and secondary structure (helix, strand and coil). Therefore, dividing amino acids into three groups appears to be a more rational choice as have been used by a number of studies (2–6,8).

The ranges of these numerical values and the amino acids belonging to each group are shown in Table 1. Three descriptors, composition (C), transition (T) and distribution (D), are then computed for a given attribute to describe the global percent composition of each of the three groups of amino acids in a protein, the percent frequencies with which the attribute changes its index along the entire length of the protein, and the distribution pattern of the attribute along the sequence, respectively.

Computation of these features can be illustrated by using hydrophobicity attribute as an example. All amino acids are divided into three groups: polar, neutral and hydrophobic. The composition descriptor C consists of three values: the global percent compositions of polar, neutral and hydrophobic residues in the protein. The transition descriptor T also consists of three values: the percent frequency with which a polar residue is followed by a neutral residue or a neutral residue by a polar residue, a polar residue is followed by a hydrophobic residue or a hydrophobic residue by a polar residue, and a neutral residue is followed by a hydrophobic residue or a hydrophobic residue by a neutral residue. The distribution descriptor D consists of five values for each of the three groups: the fractions of the entire sequence, where the first residue of a given group is located, and where 25, 50, 75 and 100% of those are contained. There are 3 descriptors and $3(C) + 3(T) + 5 \times 3(D) = 21$ descriptor values for the hydrophobicity attribute. Consequently, the seven different amino acid attributes produce a total of $7 \times 3 = 21$ descriptors and $7 \times 21 = 147$ descriptor values (Table 2).

The sequence-order features can also be used for representing amino acid distribution patterns of a specific physicochemical property along a protein or peptide sequence (11,27), which have been used for predicting protein subcellular locations at accuracy levels of 72.5–88.9% (9,10). These descriptors are derived from both the Schneider–Wrede physicochemical distance matrix (9–11) and the Grantham chemical distance matrix (27) between each pair of the 20 amino acids. The dth rank sequence-order-coupling number is defined as:

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2 \quad d = 1, 2, \ldots, 30 \qquad \mathbf{5}$$

where $d_{i,i+d}$ is the distance between the two amino acids at position i and $i+d$.

For each amino acid type, the type-1 quasi-sequence-order descriptor can be defined as:

$$Xr = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \tau_d} \quad r = 1, 2, 3, \ldots, 20 \qquad \mathbf{6}$$

where f_r is the normalized occurrence of amino acid type i and w is a weighting factor($w = 0.1$). The type-2 quasi-sequence-order is defined as:

$$Xd = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{30} \tau_d} \quad d = 21, 22, 23, \ldots 50 \qquad \mathbf{7}$$

The PROFEAT implementation of each of these algorithms was extensively tested by using a number of test sequences, such as homopolymers and copolymers of different types of amino acids. The computed descriptor values were compared to the known values for these sequences to ensure that they match with each other.

DISCUSSION

The usefulness of the features covered by PROFEAT for computing the structural and physicochemical features of proteins and peptides has been tested by a number of published studies of the development of support vector machine (SVM) classification systems for predicting protein functional classes (4,26), protein–protein interactions (8) and MHC-binding peptides (J. Cui, L. Y. Han, H. H. Lin, H. L. Zhang, Z. Q. Tang, C. J. Zheng, Z. W. Cao and Y. Z. Chen, manuscript submitted). These SVM classification systems have been found to give prediction performance with sensitivity and specificity in the range of 53.0–99.3% and 82.1–99.9%, respectively. Because of the use of these structural and physicochemical features, these SVM classification systems do not rely on sequence similarity, clustering or profiles for predicting protein functional classes, and they have been found to be particularly useful for facilitating the prediction of novel proteins (13,41,42).

Moreover, the predicted descriptors important for specific classes of proteins have been found to correlate with the

Table 2. Amino acid attributes and the division of the amino acids into three groups for each attribute

Attribute	Divisions		
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R,K,E,D,Q,N	G, A, S,T,P,H,Y	C,L,V,I,M,F,W
Normalized van der Waals volume	Volume range 0–2.78	Volume range 2.95–94.0	Volume range 4.03–8.08
	G,A,S,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W
Polarity	Polarity value 4.9–6.2	Polarity value 8.0–9.2	Polarity value 10.4–13.0
	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Polarizability	Polarizability value 0–1.08	Polarizability value 0.128–120.186	Polarizability value 0.219–0.409
	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W
Charge	Positive	Neutral	Negative
	KR	ANCQGHILMFPSTWYV	DE
Secondary structure	Helix	Strand	Coil
	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	Buried	Exposed	Intermediate
	ALFCGIVW	PKQEND	MPSTHY

The division is based on the clusters of the amino acid indices of Tomii and Kanehisa (5,40) for each of the seven attributes. For such attributes as secondary structure and solvent accessibility, the division is based on statistical appearance of each amino acid in a specific state.

experimentally estimated interactions and forces that define the distinguished activities of these proteins (4,26,43). For instance, an analysis the SVM prediction of transporters have shown that, in order of prominence, hydrophobicity, amino acid composition, polarity and charge play prominent roles for identifying transporters (26). Amino acid composition and hydrophobicity are important factors for the interaction of a protein with other biomolecules. Studies of structure-activity relationships of transporter-substrate binding has shown that hydrophobic contact, hydrogen bonding (which arises primarily from polar interaction) and charged center play important roles in substrate binding (44,45). Molecular modeling have also shown that hydrophobic contact and hydrogen bonding plays important role in transporter-substrate binding (46). A new SVM prediction system was developed for predicting members and non-members of three separate transporter families TC1.C, TC3.E and TC9.A by using this reduced set of descriptors and the same protein datasets as those of the earlier study of SVM prediction of transporters that used a full set of group 5 descriptors (46), which gives a similar prediction performance, suggesting that the selected descriptors are highly useful for distinguishing members and non-members of these transporter families.

So far, individual group of features has been separately used for computing structural, functional and interaction profiles of proteins and peptides. For instance, a protein functional class prediction server SVMProt has been developed by using descriptors of the fifth feature group (2). It is of interest to examine how the use of additional features affects the performance of this and other prediction systems. For such a purpose, two new SVM systems were developed for predicting members and non-members of the enzyme EC1.15 family and transporter TC2.C family, respectively by using descriptors of all six feature groups and the same datasets as those used for developing the corresponding SVMProt prediction systems (3,26). Comparison of the results of these new SVM systems with those of the corresponding SVMProt systems shows that the sensitivity (percentage of correctly predicted family members) is increased from 92.5 to 94.4% for the EC1.15 and from 76.5 to 83.2% for the TC2.C family, respectively, while the specificity (percentage of correctly predicted non-family members) remains unchanged at 99.8% for both families. This seems to suggest that at least for some protein families the use of additional features can moderately improve the performance of SVMProt. The contribution of each of these features can be estimated by separately conducting SVM classification using each feature (47,48). By using the same method, the order of contribution from each of the feature groups was found to be: 5th-group (composition, transition and distribution) > 1st-group (amino acid and dipeptide composition) > 6th group (sequence-order) > autocorrelation 1 > autocorrelation 3 > autocorrelation 2. Investigation of other prediction systems and on more extensive range of protein structural, functional and interaction profiles is warranted.

The commonly-used structural and physicochemical features appear to be useful in the development of statistical learning systems for predicting protein structural classes (19,20), functional families (2–5,16,26), protein–protein interactions (6,8), subcellular locations (9,10,17,18) and peptides of specific properties (J. Cui, L. Y. Han, H. H. Lin, H. L. Zhang, Z. Q. Tang, C. J. Zheng, Z. W. Cao and Y. Z. Chen, manuscript submitted). Various proteins are known to form covalent bonding with their substrates and inhibitors. These types of properties are unlikely to be sufficiently covered by the existing set of features. Some of the molecular descriptors widely used in describing the structural and physicochemical properties of chemical compounds (49–52) may be extended for representing these features. PROFEAT can be further improved by allowing the input of new structural and physicochemical properties, expanding the program for computing additional descriptors, and providing user-friendly facilities to feed computed features into the general and specialized SVM-based servers such as GIST and SVMProt.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by Singapore ARF R148-000-072-112.

Conflict of interest statement. None declared.

REFERENCES

1. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
2. Cai,C.Z., Han,L.Y., Ji,Z.L., Chen,X. and Chen,Y.Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692–3697.
3. Cai,C.Z., Han,L.Y., Ji,Z.L. and Chen,Y.Z. (2004) Enzyme family classification by support vector machines. *Proteins*, **55**, 66–76.
4. Han,L.Y., Cai,C.Z., Lo,S.L., Chung,M.C. and Chen,Y.Z. (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, **10**, 355–368.
5. Dubchak,I., Muchnik,I., Mayor,C., Dralyuk,I. and Kim,S.H. (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*, **35**, 401–407.
6. Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.
7. Bock,J.R. and Gough,D.A. (2003) Whole-proteome interaction mining. *Bioinformatics*, **19**, 125–134.
8. Lo,S.L., Cai,C.Z., Chen,Y.Z. and Chung,M.C. (2005) Effect of training datasets on support vector machine prediction of protein–protein interactions. *Proteomics*, **5**, 876–884.
9. Chou,K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **278**, 477–483.
10. Chou,K.C. and Cai,Y.D. (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem. Biophys. Res. Commun.*, **320**, 1236–1239.
11. Schneider,G. and Wrede,P. (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: *de novo* design of an idealized leader peptidase cleavage site. *Biophys. J.*, **66**, 335–344.
12. Han,L.Y., Cai,C.Z., Ji,Z.L., Cao,Z.W., Cui,J. and Chen,Y.Z. (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.*, **32**, 6437–6444.
13. Gasteiger,E., Hoogland,C., Gattiker,A., Duvaud,S., Wilkins,M.R., Appel,R.D. and Bairoch,A. (2005) In Walker,J.M. (ed.), *The Proteomics Protocols Handbook*. Humana Press, Hatfield, pp. 571–607.
14. Pavlidis,P., Wapinski,I. and Noble,W.S. (2004) Support vector machine classification on the web. *Bioinformatics*, **20**, 586–587.
15. Shepherd,A.J., Gorse,D. and Thornton,J.M. (2003) A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks. *Proteins*, **50**, 290–302.

16. Bhasin,M. and Raghava,G.P. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.*, **279**, 23262–23266.
17. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
18. Chou,K.C. and Cai,Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **277**, 45765–45769.
19. Grassmann,J., Reczko,M., Suhai,S. and Edler,L. (1999) Protein fold class prediction: new methods of statistical classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 106–112.
20. Reczko,M. and Bohr,H. (1994) The DEF data base of sequence based protein fold class predictions. *Nucleic Acids Res.*, **22**, 3616–3619.
21. Feng,Z.P. and Zhang,C.T. (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.*, **19**, 269–275.
22. Lin,Z. and Pan,X.M. (2001) Accurate prediction of protein secondary structural content. *J. Protein Chem.*, **20**, 217–220.
23. Horne,D.S. (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers*, **27**, 451–477.
24. Sokal,R.R. and Thomson,B.A. (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am. J. Phys. Anthropol.*, **129**, 121–131.
25. Dubchak,I., Muchnik,I., Holbrook,S.R. and Kim,S.H. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad. Sci. USA*, **92**, 8700–8704.
26. Lin,H.H., Han,L.Y., Cai,C.Z., Ji,Z.L. and Chen,Y.Z. (2006) Prediction of transporter family from protein sequence by support vector machine approach. *Proteins*, **62**, 218–231.
27. Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
28. Broto,P., Moreau,G. and Vandicke,C. (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.*, **19**, 71–78.
29. Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.
30. Cid,H., Bunster,M., Canales,M. and Gazitua,F. (1992) Hydrophobicity and structural classes in proteins. *Protein Eng.*, **5**, 373–375.
31. Bhaskaran,R. and Ponnuswammy,P.K. (1988) Positional flexibilities of amino acid residues in globular proteins. *Int. J. Pept. Protein Res.*, **32**, 242–255.
32. Charton,M. and Charton,B.I. (1982) The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.*, **99**, 629–644.
33. Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1–12.
34. Bigelow,C.C. (1967) On the average hydrophobicity of proteins and the relation between it and protein structure. *J. Theor. Biol.*, **16**, 187–211.
35. Charton,M. (1981) Protein folding and the genetic code: an alternative quantitative model. *J. Theor. Biol.*, **91**, 115–123.
36. Dayhoff,H. and Calderone,H. (1978) Composition of Proteins. *Altas of Protein Sequence and Structure*, **5**, 363–373.
37. Moreau,G. and Broto,P. (1980) Autocorrelation of molecular structures, application to SAR studies. *Nour. J. Chim.*, **4**, 757–764.
38. Moran,P.A. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.
39. Geary,R.C. (1954) The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**, 115–145.
40. Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.
41. Cui,J., Han,L.Y., Cai,C.Z., Zheng,C.J., Ji,Z.L. and Chen,Y.Z. (2005) Prediction of functional class of novel bacterial proteins without the use of sequence similarity by a statistical learning method. *J. Mol. Microbiol. Biotechnol*, **9**, 86–100.
42. Han,L.Y., Cai,C.Z., Ji,Z.L. and Chen,Y.Z. (2005) Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology*, **331**, 136–143.
43. Lin,H.H., Han,L.Y., Zhang,H.L., Zheng,C.J., Xie,B. and Chen,Y.Z. (2006) Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J. Lipid Res.*, **47**, 824–831.
44. Schleifer,K.J. and Tot,E. (1999) Molecular modeling study of diltiazem mimics at L-type calcium channels. *Pharm. Res.*, **16**, 1506–1513.
45. Seelig,A. and Landwojtowicz,E. (2000) Structure-activity relationship of P-glycoprotein substrates and modifiers. *Eur. J. Pharm. Sci.*, **12**, 31–40.
46. Fu,W., Cui,M., Briggs,J.M., Huang,X., Xiong,B., Zhang,Y., Luo,X., Shen,J., Ji,R., Jiang,H. *et al.* (2002) Brownian dynamics simulations of the recognition of the scorpion toxin maurotoxin with the voltage-gated potassium ion channels. *Biophys. J.*, **83**, 2370–2385.
47. Draper,D.E. (1999) Themes in RNA-protein recognition. *J. Mol. Biol.*, **293**, 255–270.
48. Ding,C.H. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.
49. Todeschini,R. and Consonni,V. (2000) *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, Germany, pp. 667.
50. Hall,L.H., Kellogg,G.E. and Haney,D.N. (2002) *Molconn-Z 4.00 User's Guide*. Edusoft-lc, Inc, Ashland, VA.
51. Wegner,J.K. and Zell,A. (2002) *JOELib-a java based computational chemistry package*. 16th Molecular Modeling Workshop, Dermstadt.
52. Xue,Y., Li,Z.R., Yap,C.W., Sun,L.Z., Chen,X. and Chen,Y.Z. (2004) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inform. Comp. Sci.*, **44**, 1630–1638.